

An adaptable generalization of Hotelling's T^2 test in high dimension

Haoran Li* Alexander Aue* Debashis Paul* Jie Peng* Pei Wang†

September 29, 2016

Abstract

In this paper, we propose a test for the two-sample problem of testing equality of mean vectors in the high-dimensional regime. The proposed test is based on a ridge-regularized Hotelling's T^2 statistic. We derive the cut-off values for the test through asymptotic analysis and suggest several finite sample modifications. We also propose a data driven choice of the regularization parameter as well as a composite testing procedure aiming at maximizing power under a class of local alternatives. Although the test is derived under Gaussianity, extensions to a class of non-Gaussian observations are also established. Through an extensive simulation study, the test is shown to compare favorably against a host of existing methods designed to tackle high-dimensional testing problems in a wide range of settings. Finally, the proposed methodology is applied to a breast cancer data set from the Cancer Genome Atlas.

Keywords: Covariance matrix, Hotelling's T^2 statistic, hypothesis testing, locally most powerful tests, random matrix theory.

AMS Subject Classification: 62H15, 62E10.

1 Introduction

Inference on high-dimensional data has remained a central topic of statistical research for over a decade. During this period, remarkable progress has been made through the use of a few well-studied principles such as sparsity and convex penalization in various contexts, including variable selection in linear models, estimation of graphical models, estimation of covariance matrices, classification problems and hypothesis testing problems. The focus of this paper is on the classical problem of testing for the equality of means of two populations having an unknown but equal covariance matrix, when the dimension is comparable to the sample size. The proposed methodology is first formulated for normal populations, and then extensions to non-Gaussian settings are established.

The standard solution to the two-sample testing problem is the well-known Hotelling's T^2 test (Anderson, 1984; Muirhead, 1982). In spite of its central role in classical multivariate statistics, Hotelling's T^2 test has several limitations when dealing with data whose dimension p is comparable to, or larger than, the sample

*Department of Statistics, University of California at Davis, One Shields Ave, Davis, CA 95616, USA, email: [hrli, aaue, debpaul, jiepeng]@ucdavis.edu

†Icahn Institute and Department of Genetics and Genomic Sciences, School of Medicine at Mount Sinai, 1470 Madison Avenue, New York, NY 10029, USA, email: pei.wang@mssm.edu

sizes. First, due to the singularity of the sample covariance matrix, the test statistic is not defined for $p \geq n = n_1 + n_2$, where n_1 and n_2 denote the sizes for the two samples. Even for $p < n$ such that the ratio p/n is close to 1, the test is known to perform badly. The precise behavior of the test statistic, in a framework where p and n grow simultaneously, so that $p/n \rightarrow \gamma \in (0, 1)$, was studied by Bai & Saranadasa (1996), who proved that in this setting the test is inconsistent.

Many approaches have been proposed in the literature to correct for the inconsistency of Hotelling's T^2 in high dimensions. Bai & Saranadasa (1996) proposed a statistic based on the squared Euclidean norm $\|\bar{X}_1 - \bar{X}_2\|^2$, where $\{X_{11}, \dots, X_{1n_1}\}$ and $\{X_{21}, \dots, X_{2n_2}\}$ denote the two samples, thus bypassing the need for estimating the population covariance Σ . Srivastava & Du (2008) and Srivastava (2009) replaced the inverse of the sample covariance matrix with the inverse of the diagonal of the sample covariance matrix. Dong et al. (2016) modified the procedure of Srivastava & Du (2008) by using shrinkage-based estimates of the variances, which was shown to perform better when $p > n$. Chen & Qin (2010) modified the procedure proposed by Bai & Saranadasa (1996) by removing the cross-product terms in $\|\bar{X}_1 - \bar{X}_2\|^2$, thus enabling consistency of the test even when the dimension p is much larger than the sample size n . Wang et al. (2015) developed a modified version of the test by Chen & Qin (2010) in a one sample setting by means of spatial sign functions, and derived the asymptotic distribution of the test statistic under a class of elliptically symmetric distributions. Chen et al. (2011) considered the one-sample testing problem and modified the inverse sample covariance S^{-1} appearing in Hotelling's T^2 to $(S + \lambda I_p)^{-1}$, where $\lambda \geq 0$ is a regularization parameter chosen in a data-dependent manner. In related developments, Wang et al. (2013) proposed a jackknife empirical likelihood test for the equality of means in moderately high dimensions. These test statistics are all based on estimators of $\|A(\mu_1 - \mu_2)\|^2$ for some given positive definite matrix A . Lopes et al. (2011) took a different approach, considering random projections of the data into a certain low-dimensional space and averaging Hotelling's T^2 computed from the projected data. Recently, Srivastava et al. (2016) adopted a closely related approach based on multiple random projections. Moreover, Biswas & Ghosh (2014) considered a nonparametric, graph-based two-sample test in high-dimensional settings.

All the above methods are designed for what may be referred to as a “dense alternative” setting, meaning that the mean vector μ in the one-sample problem, or the difference of mean vectors $\mu_1 - \mu_2$ in the two-sample problem are not assumed to have any particular structure. A different approach is taken by Cai et al. (2014) who assumed that the mean vector (or their difference in the two-sample case) is sparse. In addition, assuming that a “good” estimate of the precision matrix $\Omega = \Sigma^{-1}$ is available, they constructed a test statistic as the maximum of the squared two-sample t -statistics of the transformed observations $\{\Omega X_{11}, \dots, \Omega X_{1n_1}\}$ and $\{\Omega X_{21}, \dots, \Omega X_{2n_2}\}$. Among other recent contributions to the literature, Wang et al. (2015) developed nonparametric tests for high-dimensional mean vectors, Gregory et al. (2015) designed tests for the “large p , small n ” situation, Chen, Li & Zhong (2014) proposed a test based on a combination of data transformation via the precision matrix together with a thresholding of coordinates, Chang, Zhou & Zhou (2014) considered a maximum of t -statistics under unequal variance assumption and employed a simulation-based cut-off to detect sparse alternatives, and Guo & Chen (2016) studied hypothesis testing under high-dimensional generalized linear models.

In this paper, we generalize the *Regularized Hotelling's T^2 (RHT)* procedure proposed by Chen et al. (2011) for one-sample tests to both one- and two-sample settings, though for simplicity of exposition, all the results are presented in the two-sample case. The key contribution of this paper is a systematic investigation of the choice of the regularization parameter λ from the point-of-view of maximizing the power of the test. Throughout, we work under the setting where $p/n \rightarrow c \in (0, \infty)$ while the ratio of sample sizes n_1/n_2 converges to a non-zero, finite constant. We first demonstrate that a data-driven choice of λ that leads to asymptotically optimal power for certain classes of local alternatives is possible. Secondly, we propose a new composite test by combining the RHT statistics corresponding to a set of optimally chosen regularization parameters. The proposed composite RHT test procedure is easy to implement. Moreover, through an extensive simulation study, we demonstrate it to be robust in terms of having good power against a host of alternatives and over a wide range of covariance structures. In terms of theoretical properties, the test is shown to have identical asymptotic behavior for a class of sub-Gaussian distributions. Establishing the latter result is non-trivial due to the lack of independence between the sample mean and sample covariance matrices in non-Gaussian settings. Because of these properties, and since the prefixes “robust” and “adaptive” are already part of the statistical nomenclature tied to specific contexts, the new procedure is termed “adaptable RHT”, abbreviated as ARHT. Finally, we show that a simple monotone transformation of the test statistic, or a χ^2 approximation, can significantly enhance the finite-sample behavior of the proposed tests in terms of improving the accuracy of the actual level of significance without sacrificing power.

The rest of the paper is organized as follows. Section 2 introduces the RHT statistic as proposed in Chen et al. (2011) and lists some of its properties. Then, a class of local alternatives is defined with respect to which the optimal regularization parameter λ can be selected in a data-dependent manner. The adaptable RHT test statistic is defined in Section 3 and its practical implementation is discussed. Section 4 contains the finite-sample adjustments to better calibrate the Type I error of the proposed tests. Extensions to the non-Gaussian case are given in Section 5. A simulation study is reported in Section 6. An application of the proposed tests to a breast cancer data set from the Cancer Genome Atlas is described in Section 7. An overall summary and future research directions are outlined in Section 8. Technical details and further simulation results are collected in the Appendix. An on-line supplementary material is available at anson.ucdavis.edu/~lihaoran.

2 The regularized Hotelling's T^2 test

2.1 Two-sample RHT

This section introduces the two-sample regularized Hotelling's T^2 (RHT) statistic which is a generalization of the one-sample RHT statistic proposed in Chen et al. (2011). Throughout it is assumed that $X_{ij} \sim N(\mu_i, \Sigma)$, $j = 1, \dots, n_i$, $i = 1, 2$, are two independent samples with $\Sigma = \Sigma_p$ being a $p \times p$ non-negative definite matrix. Note that Gaussianity is not essential for the proposed tests and that the more general case will be treated in Section 5. The covariance matrix Σ_p can be estimated by its empirical counterpart $S_n = (n - 2)^{-1} \sum_{i=1}^2 \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)(X_{ij} - \bar{X}_i)^T$, where $n = n_1 + n_2$, \bar{X}_i is the sample mean of the i th sample, and T is used to denote transposition of matrices and vectors.

Due to the inadequacy of S_n when $p > n$, it is proposed here to formulate tests of the null hypothesis $H_0: \mu_1 = \mu_2$ based on the family of ridge-regularized Hotelling's T^2 statistics:

$$\text{RHT}(\lambda) = \frac{n_1 n_2}{n_1 + n_2} (\bar{X}_1 - \bar{X}_2)^T (S_n + \lambda I_p)^{-1} (\bar{X}_1 - \bar{X}_2), \quad (2.1)$$

indexed by a tuning parameter $\lambda > 0$ controlling the regularization strength. Observe that taking λ to infinity leads to the procedure of Bai & Saranadasa (1996). The limiting behavior of $\text{RHT}(\lambda)$ is tied to the spectral properties of Σ_p . Denote by $\tau_{1,p} \geq \dots \geq \tau_{p,p} \geq 0$ the eigenvalues of Σ_p , and by $H_p(\tau) = p^{-1} \sum_{\ell=1}^p \mathbf{1}_{[\tau_{\ell,p}, \infty)}(\tau)$ its empirical spectral distribution (ESD). The following conditions, which are standard assumptions in random matrix theory (RMT) literature, are imposed.

Assumption 2.1. **A1** The $p \times p$ covariance matrix Σ_p is non-negative definite and $\limsup_p \tau_{1,p} < \infty$;

A2 (High-dimensional setting) Dimensionality p and sample size $n = n_1 + n_2$ satisfy the relations $p, n \rightarrow \infty$ such that $n_1/n \rightarrow \kappa \in (0, 1)$, $\gamma_n = p/n \rightarrow \gamma \in (0, \infty)$ and $\sqrt{n}|p/n - \gamma| \rightarrow 0$;

A3 (Stabilizing of ESD) The ESD $H_p(\tau)$ of Σ_p converges as $p \rightarrow \infty$ to a probability distribution function $H(\tau)$ at every point of continuity of H , and H is nondegenerate at 0. Moreover, $\sqrt{n}\|H_p - H\|_\infty \rightarrow 0$.

Since $\lambda > 0$ and in view of (2.1), it suffices in **A1** to require non-negative definiteness of Σ_p rather than positive definiteness. The condition $\limsup_p \tau < \infty$ is necessary to obtain eigenvalue bounds. Condition **A2** ensures a well-balanced design and defines the asymptotic regime in a way that dimensionality p and sample sizes n_1 and n_2 grow proportionately. Finally, **A3** restricts the variability allowed in H_p as p increases.

If the requirements of Assumption 2.1 are met, then the asymptotic mean and variance of $\text{RHT}(\lambda)$ can be calculated as follows. Let I_p be the $p \times p$ identity matrix and, for $z \in \mathbb{C}$, denote by $R_n(z) = (S_n - zI_p)^{-1}$ and $m_{F_{n,p}}(z) = p^{-1} \text{tr}[R_n(z)]$ the resolvent and *Stieltjes transform* of the *empirical spectral distribution (ESD)* of the sample covariance matrix S_n . For the properties of the Stieltjes transform and extensive results on the limiting behavior of the ESD of a sample covariance matrix, one may refer to Bai & Silverstein (2010). Note that, under Assumption 2.1, $m_{F_{n,p}}(z)$ converges pointwise almost surely on $\mathbb{C}_+ = \{z = u + iv : v > 0\}$ to a non-random limiting distribution with Stieltjes transform $m_F(z)$ given as the solution to the equation $m_F(z) = \int [\tau \{1 - \gamma - \gamma z m_F(z)\} - z]^{-1} dH(\tau)$. Similarly to the case of the one-sample RHT discussed in Chen et al. (2011), the asymptotic mean and variance of the two-sample $\text{RHT}(\lambda)$ under Gaussianity are given by (upto multiplicative constants):

$$\Theta_1(\lambda, \gamma) = \frac{1 - \lambda m_F(-\lambda)}{1 - \gamma \{1 - \lambda m_F(-\lambda)\}}, \quad (2.2)$$

$$\Theta_2(\lambda, \gamma) = \frac{1 - \lambda m_F(-\lambda)}{[1 - \gamma \{1 - \lambda m_F(-\lambda)\}]^3} - \lambda \frac{\{m_F(-\lambda) - \lambda m'_F(-\lambda)\}}{[1 - \gamma \{1 - \lambda m_F(-\lambda)\}]^4}. \quad (2.3)$$

Moreover, as in Chen et al. (2011), we can establish the asymptotic normality of $\text{RHT}(\lambda)$. The proofs in the two-sample case require only minor modifications under Gaussianity due to the fact that sample means are normally distributed and are independent of the sample covariances. Let ξ_α be the $1 - \alpha$ quantile of the standard normal distribution $N(0, 1)$ and replace $\Theta_j(\lambda, \gamma)$ with empirical versions $\hat{\Theta}_j(\lambda, \gamma_n)$, $j = 1, 2$, by substituting $m_F(-\lambda)$ with $m_{F_{n,p}}(-\lambda)$ and $m'_F(-\lambda)$ with $m'_{F_{n,p}}(-\lambda)$ where $m'_{F_{n,p}}(-\lambda) = p^{-1} \text{tr}[R_n^2(-\lambda)]$.

Since $\hat{\Theta}_j(\lambda, \gamma_n)$ are consistent estimators for $\Theta_j(\lambda, \gamma)$, $j = 1, 2$ (indeed, $\sqrt{p}|\hat{\Theta}_1(\lambda, \gamma_n) - \Theta_1(\lambda, \gamma)| = o_P(1)$), the RHT test rejects the null hypothesis of equal means at asymptotic level $\alpha \in (0, 1)$ if

$$T_{n,p}(\lambda) = \frac{\sqrt{p}\{p^{-1}\text{RHT}(\lambda) - \hat{\Theta}_1(\lambda, \gamma_n)\}}{\sqrt{2\hat{\Theta}_2(\lambda, \gamma_n)}} > \xi_\alpha. \quad (2.4)$$

The key to establish the facts stated above is to note that under Assumption 2.1, for every fixed $\lambda > 0$, the random matrix $R_n(-\lambda) = (S_n + \lambda I)^{-1}$ has a *deterministic equivalent* (Bai & Silverstein, 2010; Liu et al., 2015) given by

$$D_n(-\lambda) = \left(\frac{1}{1 + \gamma\Theta_1(\lambda, \gamma)} \Sigma_p + \lambda I_p \right)^{-1} \quad (2.5)$$

in the sense that, for a sequence of symmetric matrices A with bounded operator norm,

$$\frac{1}{p} \text{tr}[R_n(-\lambda)A] - \frac{1}{p} \text{tr}[D_n(-\lambda)A] \rightarrow 0 \quad (2.6)$$

as $n \rightarrow \infty$, where the convergence is with probability one. The connection between (2.6) and (2.2), (2.3) can then be seen from the relations

$$\begin{aligned} \Theta_1(\lambda, \gamma) &= \frac{1}{p} \text{tr}[R_n(-\lambda)\Sigma_p] + o_p\left(\frac{1}{\sqrt{n}}\right), \\ \Theta_2(\lambda, \gamma) &= \frac{1}{p} \text{tr}[R_n(-\lambda)\Sigma_p R_n(-\lambda)\Sigma_p] + o_p(1), \end{aligned} \quad (2.7)$$

which were established in Lemma 2 of Chen et al. (2011).

2.2 Asymptotic power

This section deals with the behavior of the regularized Hotelling's T^2 test $\text{RHT}(\lambda)$ under local alternatives. Defining $\mu = \mu_1 - \mu_2$, it is specifically assumed that the mean difference parameter vector is local in the sense that

$$\sqrt{n}\mu^T D_n(-\lambda)\mu \rightarrow q(\lambda, \gamma) \quad (2.8)$$

as $n \rightarrow \infty$ for some $q(\lambda, \gamma) > 0$, where $D_n(-\lambda)$ is the deterministic equivalent defined in (2.5). The following result determines the limit of the local power $\beta_n(\mu, \lambda) = \mathbb{P}_\mu(T_n(\lambda) > \xi_\alpha)$ of the $\text{RHT}(\lambda)$ test.

Theorem 2.1. *Suppose that Assumption 2.1 and the stability condition (2.8) hold. Then, for any $\lambda > 0$,*

$$\beta_n(\mu, \lambda) \rightarrow \Phi\left(-\xi_\alpha + \kappa(1 - \kappa) \frac{q(\lambda, \gamma)}{\sqrt{2\gamma\Theta_2(\lambda, \gamma)}}\right), \quad (2.9)$$

as $n \rightarrow \infty$, where Φ denotes the cumulative distribution function of $N(0, 1)$ distribution and $\Theta_2(\lambda, \gamma)$ is defined in (2.3).

Remark 2.1. (a) Let \mathbf{E}_j denote the eigen-projection matrix associated with the j th largest eigenvalue $\tau_{j,p}$ of Σ_p . A sufficient condition for (2.8) to hold is given by the existence of a sequence of continuous functions

$f_p: \mathbb{R} \rightarrow \mathbb{R}$ satisfying $f_p(\tau_{j,p}) = p\sqrt{n}\|\mathbf{E}_j\mu\|^2$, $j = 1, \dots, p$, such that f_p converges uniformly to a function f_∞ as $n, p \rightarrow \infty$. In this case,

$$\begin{aligned} q(\lambda, \gamma) &= \{1 + \gamma\Theta_1(\lambda, \gamma)\} \int \frac{f_\infty(\tau)dH(\tau)}{\tau + \lambda\{1 + \gamma\Theta_1(\lambda, \gamma)\}} \\ &= \int \frac{f_\infty(\tau)dH(\tau)}{\tau\{1 - \gamma(1 - \lambda m_F(-\lambda))\} + \lambda}. \end{aligned} \quad (2.10)$$

Here the second line in (2.10) follows from the relationship $\{1 + \gamma\Theta_1(\lambda, \gamma)\}^{-1} = 1 - \gamma + \lambda\gamma m_F(-\lambda)$, $z \in \mathbb{C}$.

(b) If $\Sigma_p = I_p$, then (2.8) is satisfied if $\sqrt{n}\|\mu\|^2 \rightarrow c^2 > 0$. In this case $q(\lambda, \gamma) = c^2\Theta_1(\lambda, \gamma)$.

(c) Let f_∞ be as defined in part (a). If $f_\infty(x) = 1$ on the support of H , then $q(\lambda, \gamma) = m_F(-\lambda)$. If $f_\infty(x) = x$ on the support of H , then $q(\lambda, \gamma) = \Theta_1(\lambda, \gamma)$.

2.3 Selecting the regularization parameter λ

This section discusses the selection of the ridge-type regularization parameter λ in a fully data-dependent manner. A potential strategy may be based on the maximization of the local asymptotic power $\beta_n(\mu, \lambda)$. Theorem 2.1 then yields that λ should be chosen such that the ratio $Q(\lambda, \gamma) = q(\lambda, \gamma)/\sqrt{\gamma\Theta_2(\lambda, \gamma)}$ is maximized. However, in general the function $q(\lambda, \gamma)$ is based on the asymptotic behavior of the unknown mean difference vector μ .

One way to circumvent the dependence on the unknown parameter μ is to make certain assumptions about the asymptotic behavior of μ , for example of the types described in parts (a) and (c) of Remark 2.1. Here, we pursue a different approach, namely, treating μ as a random vector with mean 0 and covariance matrix proportional to $\sum_{m=0}^2 \pi_m \Sigma^m$, for some pre-specified values $\pi_0, \pi_1, \pi_2 \geq 0$ with $\sum_{m=0}^2 \pi_m = 1$, i.e.,

$$\mu \sim N(0, C_\mu n^{-1/2} p^{-1} \sum_{m=0}^2 \pi_m \Sigma^m), \quad (2.11)$$

for some constant $C_\mu > 0$.

We then consider the behavior of $\mathbb{E}[\beta_n(\mu, \lambda)]$ rather than $\beta_n(\mu, \lambda)$ itself. Specifically, we select the regularization parameter λ as the value maximizing an estimate of the power function under the above model of μ . The discussion below shows that this can be done without any explicit knowledge of Σ_p .

Observe that, under (2.11), $\beta_n(\mu, \lambda)$ concentrates around $\mathbb{E}[\beta_n(\mu, \lambda)]$, since $\beta_n(\mu, \lambda)$ is a function of $\mu^T D_n(-\lambda)\mu$, which concentrates around its mean for large enough n (and p). Moreover, under (2.11)

$$q(\lambda, \gamma) = \sum_{m=0}^2 \pi_m \rho_m(-\lambda, \gamma), \quad (2.12)$$

where

$$\begin{aligned} \rho_0(-\lambda, \gamma) &= m_F(-\lambda), \\ \rho_1(-\lambda, \gamma) &= \Theta_1(\lambda, \gamma), \\ \rho_2(-\lambda, \gamma) &= \{1 + \gamma\Theta_1(\lambda, \gamma)\}\{\phi - \lambda\rho_1(-\lambda, \gamma)\}, \end{aligned}$$

with $\phi = \int \tau dH(\tau)$ denoting the mean of the limiting spectral distribution $H(\cdot)$. Note that ϕ can be estimated accurately by $\hat{\phi} = p^{-1} \text{tr}(S_n)$. This statement is made precise in Proposition A.1 in the Appendix, which determines an exponential bound for the tail probabilities of $|\hat{\phi} - \phi|$. Consequently, if Assumption 2.1 is satisfied, then $\rho_m(-\lambda, \gamma)$ ($m = 0, 1, 2$) can be estimated from the data with an estimation error of order $o_p(1/\sqrt{n})$.

Remark 2.2. More generally, for any integer j , ρ_j satisfies the recursive equation

$$\rho_{j+1}(-\lambda, \gamma) = \{1 + \gamma \Theta_1(\lambda, \gamma)\} \left\{ \int \tau^j dH(\tau) - \lambda \rho_j(-\lambda, \gamma) \right\}.$$

However, the estimation of the higher-order moments $\int \tau^j dH(\tau)$ with $j \geq 2$ is difficult. To avoid the evaluation of high-order spectral moments for the estimation of the power function, m is therefore restricted to be at most 2. This requirement is also important for going beyond normality assumptions.

The foregoing leads to the following algorithm to determine the regularization parameter λ .

Algorithm 2.1 (Empirical selection of λ). *Perform the following steps.*

- *Step 1. Choose the prior weights $\pi = (\pi_0, \pi_1, \pi_2)$;*
- *Step 2. For $j = 0, 1, 2$ and for each λ , compute estimates $\hat{\rho}_j(-\lambda, \gamma_n)$ of $\rho_j(-\lambda, \gamma)$ as follows:*
 $\hat{\rho}_0(-\lambda, \gamma_n) = m_{F_{n,p}}(-\lambda)$, $\hat{\rho}_1(-\lambda, \gamma_n) = \hat{\Theta}_1(\lambda, \gamma_n)$,
 $\hat{\rho}_2(-\lambda, \gamma_n) = \{1 + \gamma_n \hat{\Theta}_1(\lambda, \gamma_n)\} \{\hat{\phi} - \lambda \hat{\rho}_1(-\lambda, \gamma_n)\};$
- *Step 3. For each λ , compute the estimate $\hat{Q}_n(\lambda, \gamma_n; \pi) = \sum_{m=0}^2 \pi_m \hat{\rho}_m(-\lambda, \gamma_n) / (\gamma_n \hat{\Theta}_2(\lambda, \gamma_n))^{1/2}$;*
- *Step 4. Select the regularization parameter as $\lambda_\pi \equiv \lambda_{\pi,n} = \arg \max_\lambda \hat{Q}_n(\lambda, \gamma_n; \pi)$ through a grid search.*

Although in theory one can allow an arbitrarily small positive λ in the test procedure, in practice, a lower bound for λ needs to be specified to ensure stability of the test statistic when $p \approx n$ or $p > n$. In addition, for practicality of grid search, an upper bound on λ is also needed. Accordingly, the interval for λ is set to be $[\underline{\lambda}, \bar{\lambda}]$. We recommend the values $\underline{\lambda} = p^{-1} \text{tr}(S_n)/100$ and $\bar{\lambda} = 20 \|S_n\|$.

The behavior of the test with the data-driven tuning parameter is described in the following theorem.

Theorem 2.2. *Let $[\underline{\lambda}, \bar{\lambda}]$ (with $\bar{\lambda} > \underline{\lambda} > 0$) be a non-empty interval. If the conditions of Assumption 2.1 are satisfied and if there is a $C > 0$ such that $\frac{\partial^2}{\partial \lambda^2} Q(\lambda_\infty, \gamma; \pi) < -C$ for any local maximizer λ_∞ of $Q(\lambda, \gamma; \pi)$ on $[\underline{\lambda}, \bar{\lambda}]$, then there exists a sequence $(\lambda_n : n \in \mathbb{N})$ of local maximizers of $(\hat{Q}_n(\lambda, \gamma_n; \pi) : n \in \mathbb{N})$, satisfying*

$$n^{1/4} |\lambda_n - \lambda_\infty| = O_p(1) \quad (n \rightarrow \infty). \quad (2.13)$$

Further, under the null hypothesis,

$$T_{n,p}(\lambda_n) = \frac{\sqrt{p} \{p^{-1} \text{RHT}(\lambda_n) - \hat{\Theta}_1(\lambda_n, \gamma_n)\}}{\sqrt{2 \hat{\Theta}_2(\lambda_n, \gamma_n)}} \implies N(0, 1) \quad (n \rightarrow \infty), \quad (2.14)$$

where \implies denotes convergence in distribution. Moreover, if λ_∞ is a boundary point and $\frac{\partial}{\partial \lambda} Q(\lambda_\infty, \gamma; \pi) \neq 0$, then the assumption on $\frac{\partial^2}{\partial \lambda^2} Q(\lambda_\infty, \gamma; \pi)$ can be dropped.

The proof of Theorem 2.2 is given in Appendix A.

3 Adaptable RHT

The previous section describes a data-driven procedure for selecting the optimal regularization parameter λ for a given prespecified weight π . This approach is further extended in this section by constructing a composite testing procedure called adaptable RHT (ARHT) where several weight vectors π may be incorporated. Specifically, ARHT is defined as:

$$\text{ARHT}_{n,p}(\Pi) = \max_{\pi \in \Pi} T_{n,p}(\lambda_\pi), \quad (3.1)$$

where $T_{n,p}(\lambda)$ is the normalized RHT test statistic defined in (2.4), λ_π is defined in Step 4 of Algorithm 2.1, and $\Pi = \{\pi_1, \dots, \pi_k\}$ is a pre-specified set of weights ($k \geq 1$).

One option for Π consists of the three “canonical” weights $\pi = (1, 0, 0)$, $(0, 1, 0)$ and $(0, 0, 1)$. Other choices of Π can be conceived of but are not considered in this paper. An evaluation of this canonical selection procedure by simulation experiments is provided in Section 6.

Determination of the cut-off values for $\text{ARHT}_{n,p}(\Pi)$ requires knowing the asymptotic distribution of the process $T_{n,p} = (T_{n,p}(\lambda) : \lambda \in [\underline{\lambda}, \bar{\lambda}])$ under the null hypothesis of equal means. From here, the case where $\Lambda = \{\lambda_{\pi_1}, \dots, \lambda_{\pi_k}\}$ is a collection of finitely many regularization parameters can be easily derived. The following result holds.

Theorem 3.1. *If the conditions of Assumption 2.1 are satisfied, then, under the null hypothesis of equal mean,*

$$T_{n,p} \xrightarrow{d} Z \quad (n \rightarrow \infty),$$

where \xrightarrow{d} denotes weak convergence in the Skorohod space $D[\underline{\lambda}, \bar{\lambda}]$ and $Z = (Z(\lambda) : \lambda \in [\underline{\lambda}, \bar{\lambda}])$ denotes a centered Gaussian process with covariance function

$$\Gamma(\lambda, \lambda') = (1 + \gamma\Theta_1(\lambda, \gamma))(1 + \gamma\Theta_1(\lambda', \gamma)) \frac{\lambda'\Theta_1(\lambda', \gamma) - \lambda\Theta_1(\lambda, \gamma)}{(\lambda' - \lambda)\sqrt{\Theta_2(\lambda, \gamma)\Theta_2(\lambda', \gamma)}}, \quad \lambda \neq \lambda' \quad (3.2)$$

and $\Gamma(\lambda, \lambda) = 1$. In particular, for every $k \geq 1$ and every collection $\Lambda_k = \{\lambda_1, \dots, \lambda_k\} \subset [\underline{\lambda}, \bar{\lambda}]$, it holds that

$$(T_{n,p}(\lambda_1), \dots, T_{n,p}(\lambda_k))^T \implies N_k(0, \Gamma_k) \quad (n \rightarrow \infty),$$

where the limit on the right-hand side is a k -dimensional centered normal distribution with $k \times k$ covariance matrix $\Gamma_k = \Gamma(\Lambda_k)$ whose entries are given by $\Gamma(\lambda_i, \lambda_j)$, $i, j = 1, \dots, k$.

The proof of Theorem 3.1 is given in Appendix A. It shows that the statistic $\text{ARHT}_{n,p}(\Pi)$ has a non-degenerate limiting distribution under H_0 . Theorem 3.1 can be used directly to determine the cut-off values of the test by deriving analytical formulae for the quantiles of the limiting distribution. To avoid the associated complex calculations, we adopt an alternative approach, whereby a *parametric bootstrap* procedure is applied to approximate the cut-off values. Specifically, $\Gamma = \Gamma(\Lambda)$ is first estimated from the data by $\hat{\Gamma} = \hat{\Gamma}_n(\Lambda)$, and then bootstrap replicates are generated by simulating from $N_k(0, \hat{\Gamma})$, leading to an approximation of the null distribution of $\text{ARHT}_{n,p}(\Pi)$. Observe that a natural candidate for the covariance estimator is

$$\hat{\Gamma}_n(\lambda, \lambda') = (1 + \gamma_n \hat{\Theta}_1(\lambda, \gamma_n))(1 + \gamma_n \hat{\Theta}_1(\lambda', \gamma_n)) \frac{\lambda' \hat{\Theta}_1(\lambda', \gamma_n) - \lambda \hat{\Theta}_1(\lambda, \gamma_n)}{(\lambda' - \lambda) \sqrt{\hat{\Theta}_2(\lambda, \gamma_n) \hat{\Theta}_2(\lambda', \gamma_n)}}, \quad \lambda \neq \lambda'. \quad (3.3)$$

Remark 3.1. It should be noticed that $\hat{\Gamma}_n(\Lambda)$ defined through (3.3) may not be non-negative definite even though it is clearly symmetric. If such a case occurs, the negative definite estimator can be projected to its closest non-negative definite matrix simply by setting the negative eigenvalues to 0. The resulting covariance matrix estimator is called $\hat{\Gamma}_n^+(\Lambda)$. In practice, this matrix is used for generating the bootstraps samples for $\text{ARHT}_{n,p}(\Pi)$.

4 Calibration of Type-I error

Simulation studies reveal that the true size of the RHT statistics tends to be slightly inflated. This is because a normal approximation is being used to describe the fluctuations of a statistic that is essentially a quadratic form, and consequently has skewed distribution for finite samples. In this section, two remedies are proposed. The first remedy is based on a power transformation of the RHT statistics, reducing skewness by calibrating higher-order terms in the test statistics. A second remedy is to choose cut-off values of the RHT statistics based on quantiles of a normalized χ^2 distribution whose first two moments match those of the RHT.

4.1 Cube-root transformation

In principle, any power transformation may be considered, but empirically, a near-symmetry of the null distribution is obtained by a cube-root transformation of the RHT statistic. Therefore, only the cube-root transformation is discussed here, although other cases can be derived in the same fashion. An application of the δ -method yields:

$$\tilde{T}_{1/3}(\lambda) = \frac{\sqrt{p}\{[p^{-1}\text{RHT}(\lambda)]^{1/3} - \hat{\Theta}_1^{1/3}(\lambda, \gamma_n)\}}{(\sqrt{2}/3)\hat{\Theta}_2^{1/2}(\lambda, \gamma_n)/\hat{\Theta}_1^{2/3}(\lambda, \gamma_n)} \implies N(0, 1). \quad (4.1)$$

The transformed RHT statistic gives rise to cube-root transformed ARHT test statistic defined by

$$\text{ARHT}_{1/3}(\Pi) = \max_{\pi \in \Pi} \tilde{T}_{1/3}(\pi).$$

It is easy to check that statistical inference based on $\text{ARHT}_{1/3}(\Pi)$ for a set Π of k weight vectors is to be performed with the same covariance kernel $\Gamma = \Gamma(\cdot, \cdot)$ given in (3.2). We recommend $\text{ARHT}_{1/3}$ for most practical applications since it nearly symmetrizes the null distribution of the test statistic even for moderate sample sizes.

4.2 χ^2 -approximation of cut-off values

While the cube-root transformation is shown to be quite effective, a weighted chi-square approximation has often been used in the literature to approximate limiting distributions of generalized quadratic forms. We therefore formulate a different method for calibrating the size of the ARHT procedure. This involves setting the cut-off values as quantiles of the maximum of a set of scaled χ^2 distributions, i.e., random variables of the form $a\chi^2(\ell)$, where a is a normalizing constant and ℓ is the degree of freedom. For each pair of (a, ℓ) , the $a\chi^2(\ell)$ distribution is used to mimic the distribution of the RHT statistic in (2.1) for a given regularization parameter λ . The scale multipliers a and the degrees of freedom ℓ are selected in such a way that the first two moments and the covariances of the χ^2 variables match with those of the corresponding RHT test statistics,

up to a level of approximation. The details are given in the Supplementary Material. Unlike the cube-root transform of Section 4.1, this method only modifies cut-off values. Based on our simulations, the performance of this calibration method, in terms of the power curves, is similar to that of the cube-root transformation.

5 Extension to non-Gaussian distributions

The methodology introduced thus far will now be extended to a more general class of distributions beyond Gaussianity. The extension follows the approach taken in Chatterjee (2009). Following Definition 2.1 in Chatterjee (2009), we introduce the following class of probability measures obtained as smooth transformations of a standard normal random variable.

Definition 5.1. *For each $c_1, c_2 > 0$, let $\mathcal{L}(c_1, c_2)$ be the class of probability measures on the real line \mathbb{R} that arises as laws of random variables $u(Z)$, where Z is a standard normal random variable and u is a twice continuously differentiable function such that, for all $x \in \mathbb{R}$,*

$$|u'(x)| \leq c_1 \quad \text{and} \quad |u''(x)| \leq c_2. \quad (5.1)$$

Note that random variables in $\mathcal{L}(c_1, c_2)$ are sub-Gaussian and have continuous distribution, since u is a Lipschitz function with bounded Lipschitz constant. The idea behind using this class of distributions is described as follows. The first condition in (5.1) is used to control the magnitude of the variance of $u(Z)$, while the second condition is primarily for controlling the tail behavior of the statistic. This intuition is rigorously formalized in Chatterjee (2009). This approach is particularly attractive as it only requires establishing appropriate upper bounds for the operator norms of the gradient and Hessian matrices of the statistic (with respect to the variables), and matching the first two asymptotic moments. However, the calculations in our setting are non-trivial since they require a detailed analysis of the resolvent of the sample covariance matrix.

For the two-sample testing problem under consideration in this paper, let

$$X_{ij} = \mu_i + \Sigma_p^{1/2} Z_{ij}, \quad j = 1, \dots, n_i, \quad i = 1, 2, \quad (5.2)$$

be two independent random samples, where $Z_{ij} = (z_{ijk})_{k=1}^p$ are p -dimensional i.i.d. random vectors whose entries are i.i.d. random variables belonging to $\mathcal{L}(c_1, c_2)$ for some positive c_1 and c_2 , satisfying $\mathbb{E}z_{ijk} = 0$, $\mathbb{E}z_{ijk}^2 = 1$, $\mathbb{E}z_{ijk}^3 = 0$. Then the following holds.

Theorem 5.1. *All previously stated results hold if the observations X_{ij} are as in (5.2) with the z_{ijk} satisfying Definition 5.1 together with $\mathbb{E}z_{ijk} = 0$, $\mathbb{E}z_{ijk}^2 = 1$, $\mathbb{E}z_{ijk}^3 = 0$, and Σ_p satisfying Assumption 2.1.*

The proof of Theorem 5.1 is lengthy and is given in the Supplementary Material. The key component of the proof is to consider a modified version of the RHT statistic, where S_n is replaced with the un-centered matrix $\tilde{S}_n = n^{-1} \sum_{i=1}^2 \sum_{j=1}^{n_i} X_{ij} X_{ij}^T$. Defining $U_{kl}(\lambda) := \bar{X}_k^T (\tilde{S}_n + \lambda I_p)^{-1} \bar{X}_l$, for $1 \leq k, l \leq 2$, and by an appropriate use of δ -method, the asymptotic normality of RHT(λ) follows once the joint asymptotic normality of $(U_{11}(\lambda), U_{12}(\lambda), U_{22}(\lambda))$ is established. We make use of Theorem 2.2 of Chatterjee (2009) to achieve the latter result, whereby the conditions of Definition 5.1 come into play. The behavior of the power function of the RHT test under local alternatives follows analogously.

Remark 5.1. We would like to mention that Theorem 5.1 is expected to hold under even more general conditions than considered above. Indeed, in the one-sample testing problem, by making use of the analytical framework adopted by Pan & Zhou (2011), we have proved asymptotic normality of the RHT statistic when the condition imposed by Definition 5.1 is replaced by a bounded fourth moment assumption that is standard in spectral analysis of large covariance matrices. However, this derivation, apart from being heavily technical, also does not readily extend to the two-sample setting. This is connected to certain structural differences between one- and two-sample settings under non-Gaussianity. Whether such generalizations are feasible under the current context is a topic of future research.

6 Simulations

In this section, the proposed ARHT is compared by means of a simulation study with a host of popular competing methods, including the tests introduced by Bai & Saranadasa (1996) (BS), Chen & Qin (2010) (CQ), Lopes et al. (2011) (RP), and Cai et al. (2014) (CLX. $\Omega^{1/2}$ and CLX. Ω , corresponding to the two different transformation matrices $\Omega^{1/2}$ and Ω). In the following, ARHT, ARHT $_{1/3}$ and ARHT $_{\chi^2}$ denote the original, cubic-root transformed and χ^2 -approximated ARHT procedure introduced in Sections 3, 4.1 and 4.2, respectively.

6.1 Settings and results

In the simulations, the observations X_{ij} are as in (5.2), while two different distributions for z_{ijk} are considered, namely the $N(0, 1)$ distribution and the t -distribution with four degrees of freedom, $t_{(4)}$, rescaled to ensure a unit variance. For the normal case, the sample sizes are balanced, that is, $n_1 = n_2 = 50$. For the $t_{(4)}$ case, the sample sizes are $n_1 = 30$ and $n_2 = 70$. The dimension p is either 50 or 200, so that $\gamma = p/(n_1 + n_2) = 0.5$ or 2. Three models for the covariance matrix $\Sigma = \Sigma_p$ are considered:

- (i) The *identity matrix* (ID): Here $\Sigma = I_p$;
- (ii) The *sparse case* Σ_s : Here $\Sigma = \{p^{-1}\text{tr}(D)\}^{-1}D$ with a diagonal matrix D whose eigenvalues are given by $\tau_j = 0.01 + (0.1 + j)^6$, $j = 1, \dots, p$;
- (iii) The *dense case* Σ_d : Here $\Sigma = P^T \Sigma_s P$ with a unitary matrix P randomly generated from the Haar measure and resampled for each different setting.

The eigenvalues of Σ_s and Σ_d decay slowly to 0, so that no dominating leading eigenvalue exists. Under the alternative, for each p, Σ and each replicate, the mean difference vector $\mu = \mu_1 - \mu_2$ is randomly generated from one of the four models:

- (a) $\mu \sim N(0, cI_p)$; (b) $\mu \sim N(0, c\Sigma)$; (c) $\mu \sim N(0, c\Sigma^2)$;
- (d) μ is sparse with 5% randomly selected nonzero entries being either $-c$ or c with probability 1/2 each.

The parameter c is used to control the signal size. The choices in (a)–(d) respectively represent the cases that μ is uniform, slightly tilted towards the eigenvectors corresponding to large eigenvalues of Σ , heavily tilted towards the eigenvectors corresponding to large eigenvalues of Σ , and being sparse, respectively. In the

simulations, we choose $[\underline{\lambda}, \bar{\lambda}] = [0.01, 100]$ and use a grid with progressively coarser spacings for determining the optimal $\lambda_n \equiv \lambda_{\pi, n}$.

We conduct all the tests at the significance level $\alpha = 0.05$. There are two versions of each test, (i) utilizing (approximate) asymptotic cut-off values; and (ii) utilizing the size-adjusted cut-off values based on the actual null distribution computed by simulations. In this section, we report only results for the latter case. Also, we display the power graphs for the Gaussian case only. The power curves when $z_{jk} \sim t_{(4)}$ display similar characteristics, and are reported in the Supplementary Material. All empirical cut-off values, powers and sizes are calculated based on 10,000 replications. Empirical sizes for the various tests are shown in Table 1. Empirical power curves versus expected signal strength $(\sqrt{n}\mathbb{E}\|\mu\|_2^2)^{1/2}$ are shown in Figures 1–5. Note that, in some of the settings, several of the power curves nearly overlap, creating an occlusion effect. For the ease of illustration, we plot the power curves corresponding to $ARHT_{1/3}$ (recommended procedure) on the top layer.

	Σ	p	ARHT	$ARHT_{1/3}$	$ARHT_{\chi^2}$	BS	CQ	RP	$CLX.\Omega^{1/2}$	$CLX.\Omega$
Normal	ID	50	.0612	.0447	.0472	.0609	.0481	.0520	.0633	.0637
Normal	ID	200	.0568	.0473	.0493	.0561	.0508	.0490	.0754	.0757
Normal	Σ_d	50	.0854	.0489	.0606	.0695	.0470	.0485	.0970	.1101
Normal	Σ_d	200	.0917	.0601	.0705	.0622	.0486	.0503	.0833	.0971
Normal	Σ_s	50	.0877	.0492	.0603	.0688	.0468	.0508	.0613	.0615
Normal	Σ_s	200	.0938	.0596	.0707	.0645	.0487	.0503	.0773	.0773
$t_{(4)}$	ID	50	.0572	.0395	.0414	.0516	.0450	.0477	.0562	.0563
$t_{(4)}$	ID	200	.0541	.0447	.0456	.0518	.0505	.0504	.0611	.0611
$t_{(4)}$	Σ_d	50	.0836	.0473	.0582	.0659	.0468	.0485	.0815	.0906
$t_{(4)}$	Σ_d	200	.0912	.0582	.0692	.0590	.0484	.0507	.0759	.0838
$t_{(4)}$	Σ_s	50	.0812	.0451	.0559	.0634	.0449	.0481	.0512	.0512
$t_{(4)}$	Σ_s	200	.0872	.0551	.0656	.0565	.0469	.0474	.0638	.0638

Table 1: Empirical sizes of the various tests at the $\alpha = 0.05$ level.

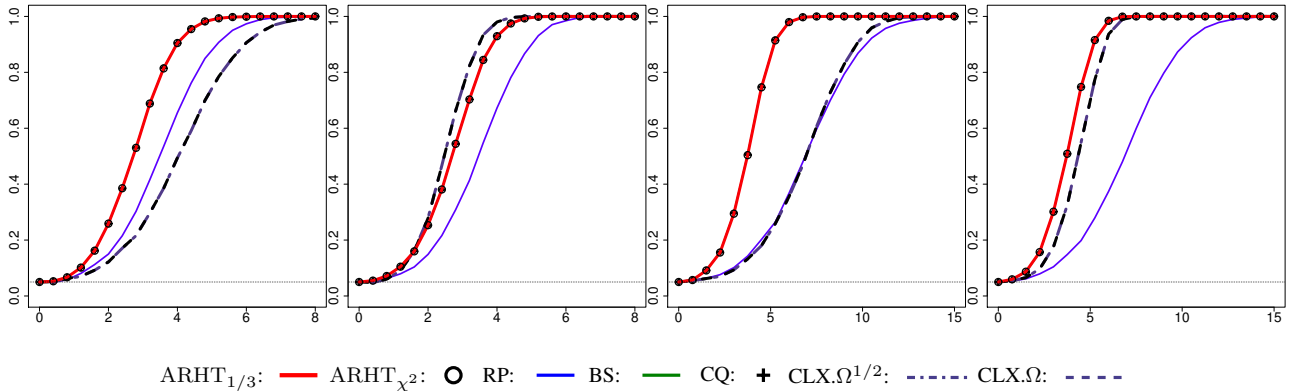


Figure 1: Size-adjusted empirical power with $X_{ij} \sim N(\cdot, \Sigma)$ and $\Sigma = \text{ID}$. From left to right: (i): $p = 50, \mu \sim N(0, cI)$; (ii): $p = 50$, sparse μ ; (iii): $p = 200, \mu \sim N(0, cI)$; (iv): $p = 200$, sparse μ .

6.2 Summary of simulation results

For each simulation configuration considered in this study, ARHT or its calibrated versions are at least comparable to the procedure(s) with the best performance, except for the case of sparse μ with relatively large p .

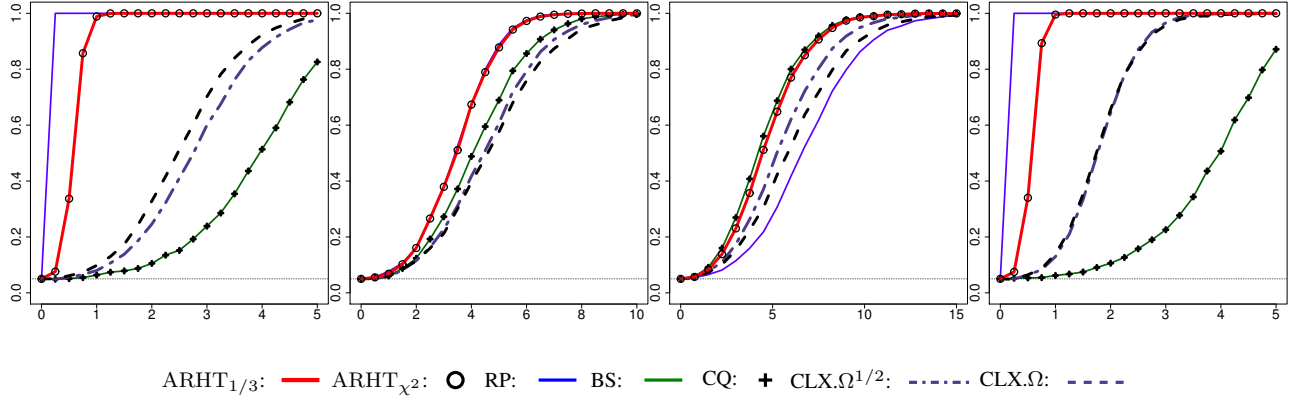


Figure 2: Size-adjusted empirical power with $X_{ij} \sim N(\cdot, \Sigma)$, $\Sigma = \Sigma_d$ and $p = 50$. From left to right: (i): $\mu \sim N(0, cI)$; (ii): $\mu \sim N(0, c\Sigma)$; (iii): $\mu \sim N(0, c\Sigma^2)$; (iv): sparse μ .

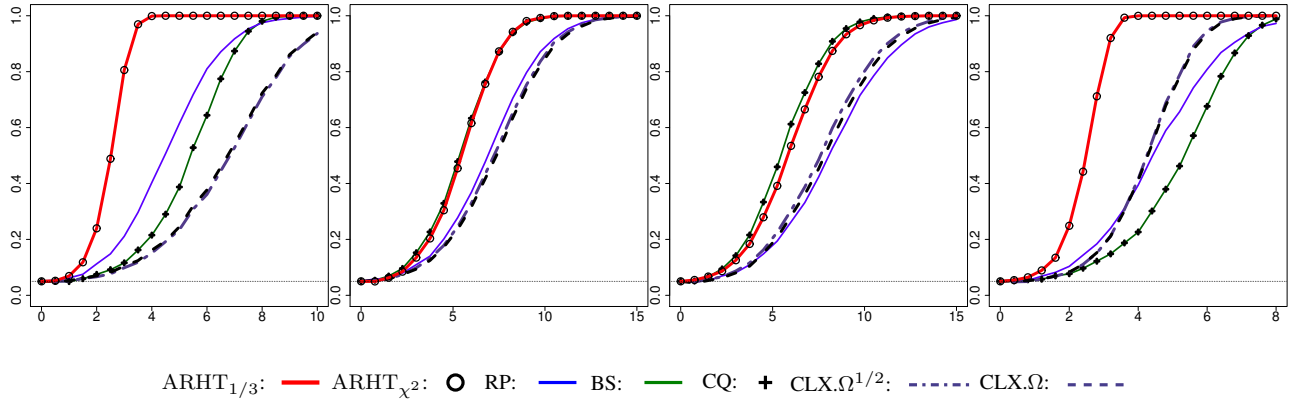


Figure 3: Same as in Figure 2, but with $p = 200$.

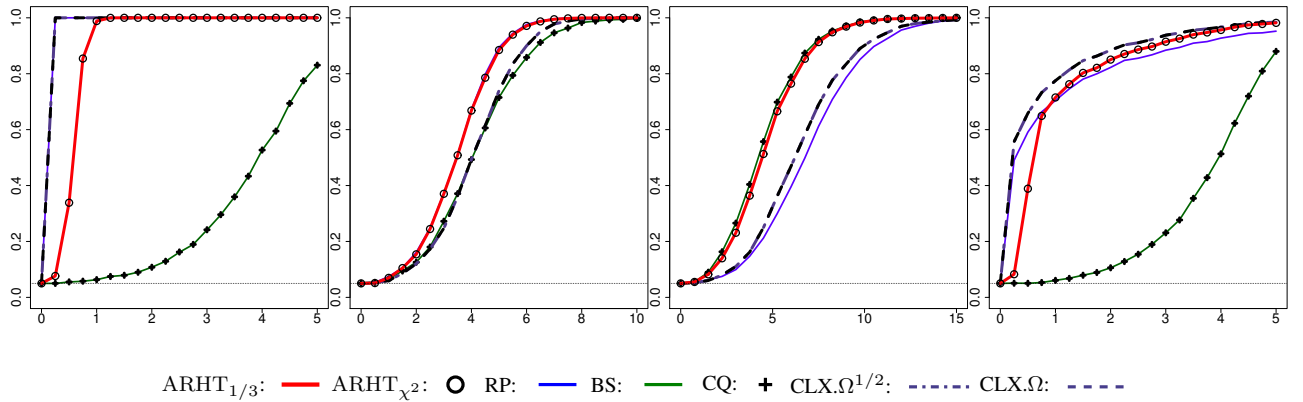


Figure 4: Size-adjusted empirical power with $X_{ij} \sim N(\cdot, \Sigma)$, $\Sigma = \Sigma_s$ and $p = 50$. From left to right: (i): $\mu \sim N(0, cI)$; (ii): $\mu \sim N(0, c\Sigma)$; (iii): $\mu \sim N(0, c\Sigma^2)$; (iv): sparse μ .

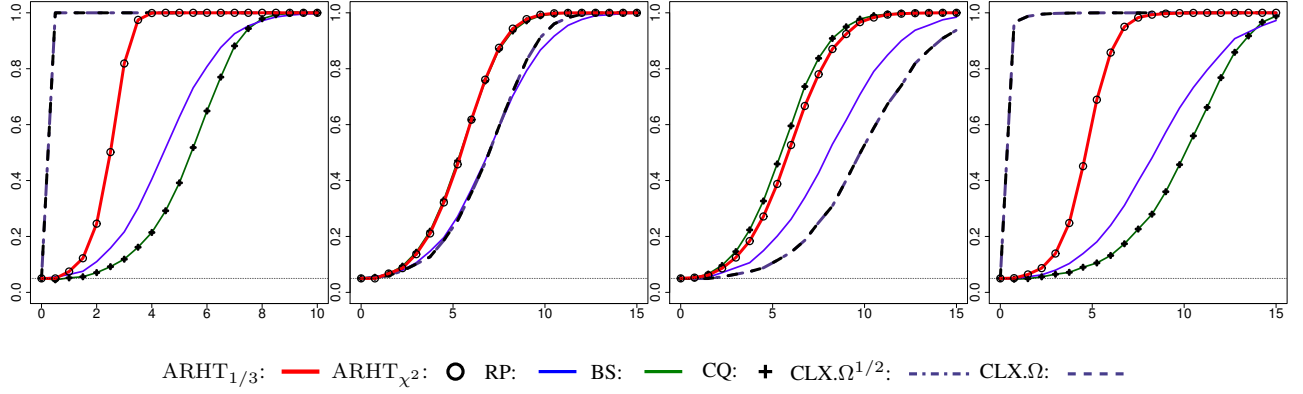


Figure 5: Same as in Figure 4, but with $p = 200$.

This serves as an evidence for the robustness of ARHT procedures with respect to the structure of the alternatives. This adaptable behavior also sets the proposed methodology apart from its competitors. The following observations are made based on the simulation outcomes.

- When the dimension is high and there is no specific structure of μ and Σ that could be exploited, ARHT tends to outperform the other tests. Tilted alternatives are expected to be detrimental to the performance of both ARHT and RP. However, ARHT can be seen as only slightly less powerful than BS and CQ, which yield the best results for this case.
- In the case that Σ is equal to the identity matrix, the BS procedure is expected to give the best performance, since the test statistic is based on the true covariance matrix. Recalling that BS can be treated as $\text{RHT}(\infty)$, ARHT is shown to perform as well as BS in corresponding simulations (see Figure 1). This may be viewed as evidence in support of the data-driven tuning parameter selection strategy detailed in Section 2.3.
- If both the mean difference vector μ and covariance matrix Σ are sparse, the three CLX procedures are expected to perform the best. Specifically, the simulations reveal that the sparsity of μ alone does not guarantee the advantage of CLX. This can be seen in the right panels of Figures 1–3. However, as evidenced in Figures 4 and 5, if Σ is sparse, then the performance of the CLX procedures is the best when μ is either uniform or sparse. The ARHT procedures are less sensitive to the structure imposed on the covariance matrix Σ than the CLX procedures, although it is less powerful in sparse settings.

The reason for the excellent performance of CLX for uniform μ (which is even better than for sparse μ) is that, significant signals occur, with high probability (due to uniform distribution of signal), at coordinates with very small variance due to their high signal-to-noise ratios. Consequently, maximum- t -statistics based methods, i.e., l_∞ -norm based methods, such as the CLX tests, are able to efficiently detect such signals. In contrast, all l_2 -norm based methods, including ARHT, combine the signals over all coordinates and thus tend to miss such signal since the l_2 norm of μ is relatively small. When μ is sparse, such a phenomenon also happens but with smaller probability. When μ is tilted, on the other hand, this phenomenon is unlikely to occur. Therefore, what is playing an important role here is not only the sparsity of μ , but also the matching of significant signals with small variance.

Results of this simulation study highlight the robustness or adaptivity of the proposed *ARHT* test to various different alternative scenarios and therefore demonstrates its potential usefulness for real world applications. In the next section its performance and that of its competitors are analyzed on a real data set.

7 Application

Breast cancer is one of the most common cancers with greater than 1,300,000 cases and 450,000 deaths each year worldwide. Breast cancer is also a heterogeneous disease, consisting of several subtypes with distinct pathological and clinical characteristics. To better understand the disease mechanisms underlying different breast cancer subtypes, it is of great interest to characterize subtype-specific somatic *copy number alteration* (CNA) patterns, which have been shown to play critical roles in activating oncogenes and in inactivating tumor suppressors during the breast tumor development; see Bergamaschi et al. (2006). In this section, we apply the proposed ARHT to a TCGA (The Cancer Genome Atlas) breast cancer data set (Cancer Genome Atlas Network, 2012) to detect pathways showing distinct CNA patterns between different breast cancer subtypes.

Level-three segmented DNA copy number (CN) data of breast cancer tumor samples were obtained from the TCGA web site. We focus on a subset of 80 breast tumor samples, which are also subjected to deep protein-profiling by CPTAC (Clinical Proteomic Tumor Analysis Consortium) (Paulovich et al., 2010; Ellis et al., 2013; Mertins et al., 2016). Thus findings from our analysis may lead to further investigations and knowledge generation through the corresponding protein profiles in the future. Specifically, among these 80 samples, 18, 29, and 33 belongs to the Her2-enriched (Her2), Luminal A (Lum A) and Luminal B (Lum B) subtypes, respectively.

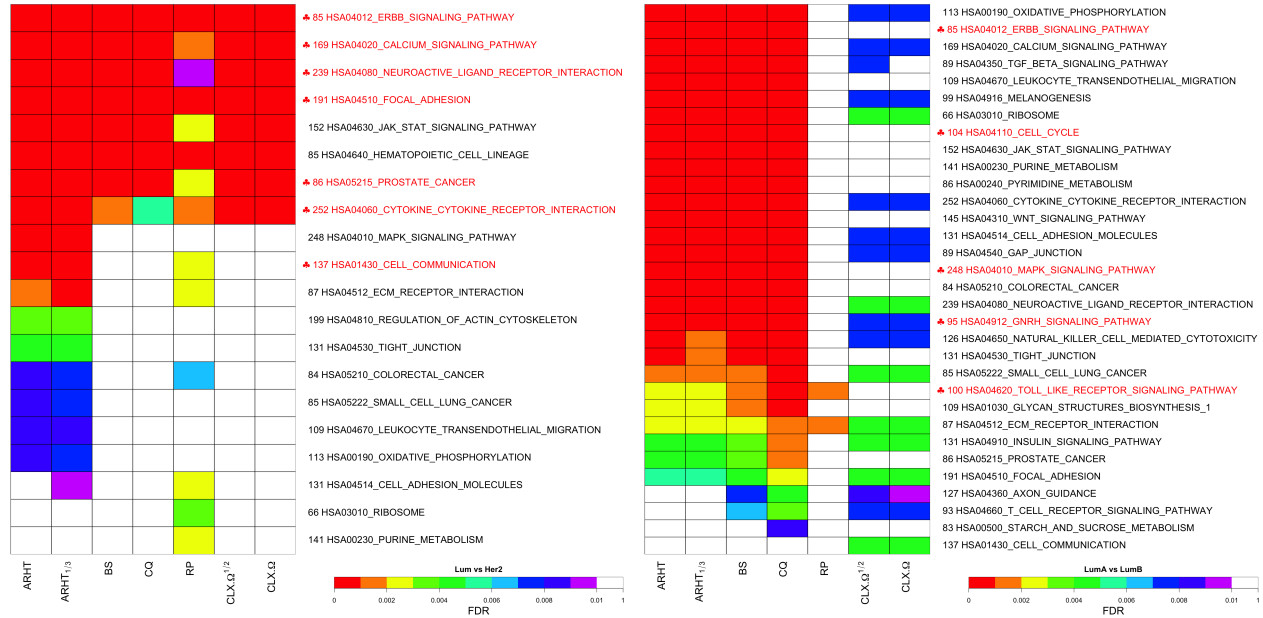


Figure 6: Lum vs Her2 (left panel) and Lum A vs Lum B (right panel). Row labels show pathway names and size (p), with those known to be significant highlighted by ♣ and red color.

For the selected samples, we first derive gene-level copy number estimates based on the segmented CN

profiles. Q-Q plots, provided in Supplementary Material, suggest that the observations have heavier tails than normal distributions. To better illustrate the comparative performance of the proposed methods under large dimensions, we consider the 36 largest KEGG pathways, with number of genes in these pathways ranging from 66 to 252, so that p/n varies between 0.75 and 3.5. For each pathway, we are interested in testing whether genes in the pathway showed different copy number alterations between Lum (Lum A plus Lum B) v.s. Her2, or Lum A vs. Lum B. These led to 72 two-sample tests in total.

We applied all 9 testing methods discussed in the simulation studies to this data set. The null distribution, and the p-value, for each method, were generated based on 100,000 permutations, instead of applying the asymptotic theory. Also, to control the family-wise error rate, the p -values are further adjusted by FDR (Benjamini & Hochberg, 1995). Note that, for m hypotheses with ordered p -values $p_{(1)} \leq \dots, p_{(m)}$, the FDR-adjusted p -values are defined as $q_{(i)} := \min_{i \leq j \leq m} (p_{(j)} m / j)$. For the Lum vs Her2 comparison, ARHT yielded the largest number of significant pathways followed by RP, while all other methods have similar behaviors and detected about half of what ARHT and RP detected. For the Lum A vs Lum B comparison, results of ARHT are similar to those of BS and CQ, giving the largest number of significant pathways. On the other hand, in this case, RP only detected two while the three CLX methods did not detect any significant pathway at FDR level of 0.01.

One unique characteristic of Her2 subtype tumors is the amplification of gene ERBB2 and its neighboring genes in cytoband 17q12, including MED1, STARD3, ect. There are 7 pathways containing at least one of these genes. These pathways, whose annotations were colored in red in Figure 6, can serve as positive controls in the Her2 vs Lum comparison (Lamy et al., 2011). Moreover, it has been shown that gene MAP3K1 and MAP2K4 have different CN loss activities in Lum A and Lum B tumors (Creighton, 2012). In addition, proliferation genes such as CCNB1, MKI67 and MYBL2 are more highly expressed in Lum B compared to Lum A, as shown in Tran and Bedard (2011). Thus, the pathways containing these genes can be viewed as positive controls in the Lum A vs Lum B comparison analysis. As an illustrative reference, in Table 2, we summarize the performance of different procedures to detect these known pathways when the FDR-adjusted significance level is 0.01. Interestingly, only the three ARHT procedures successfully detected all these pathways of positive controls, suggesting a superior power of ARHT procedures over the competitors. BS and CQ appeared to be the second best methods. These results are consistent with the observations based on the simulation study in Section 6.

Table 2: Comparative performance on known significant pathways (at FDR level 0.01).

	Lum vs Her2	Lum A vs Lum B
ARHT	7/7	5/5
ARHT _{1/3}	7/7	5/5
BS	6/7	5/5
CQ	6/7	5/5
RP	7/7	1/5
CLX. $\Omega^{1/2}$	6/7	1/5
CLX. Ω	6/7	1/5

In summary, only ARHT can consistently make correct decisions on pathways known to be significant,

while the other methods only perform well in at most one comparison. This is evidence in support of the power and robustness of ARHT.

8 Discussion

We have presented a new, computationally tractable, procedure for testing equality of means across two populations based on a composite ridge-type regularization of Hotelling's T^2 statistics. We used techniques from random matrix theory to derive the asymptotic null distribution of the statistic under a regime where the dimension is comparable to the sample sizes. We conducted extensive simulations to show that the proposed test has excellent power characteristics for a wide class of alternatives and is fairly robust to the structure of the common covariance matrix as well as the distribution of the observations. We also demonstrated the practical advantages of the proposed test in the context of a breast cancer data set where the objective was to detect DNA copy number alteration patterns in a pathway across cancer subtypes. There are several future research directions that we are pursuing. On the technical side, we aim to relax the distributional assumptions on the observations even further by only requiring the existence of a certain number of moments. On the methodological front, we aim to extend the framework to formulate tests for mean difference under possibly unequal variances, and to deal with the MANOVA problem in high-dimensional settings.

A Proof of main results

In this section, the proofs of Theorem 2.2 and Theorem 3.1 are given under the assumption of Gaussianity. The proof of Theorem 2.1 under Gaussianity can be easily obtained by making use of (2.6), and the independence between sample means and sample covariances. Hence, no details are provided here. The results stated in Section 5 under non-Gaussianity and some discussions of the extension are included in the Supplementary Material.

First, note that S_n is independent of $\bar{X}_1 - \bar{X}_2$ and has the same distribution as $\sum_{j=1}^{n-2} \Sigma_p^{1/2} Y_j Y_j^T \Sigma_p^{1/2}$ where $\sqrt{n-2} Y_j$ are i.i.d. $N(0, I_p)$ random vectors independent of the X_{ij} under Gaussianity. Redefining

$$S_n = \sum_{j=1}^{n-2} \Sigma_p^{1/2} Y_j Y_j^T \Sigma_p^{1/2}, \quad (\text{A.1})$$

therefore does not change the distribution of RHT or ARHT. For convenience, the new definition of S_n is hence adopted from now on. For any $j = 1, \dots, n-2$, define moreover

$$R_n^{(j)}(-\lambda) = (S_n - \Sigma_p^{1/2} Y_j Y_j^T \Sigma_p^{1/2} + \lambda I_p)^{-1}. \quad (\text{A.2})$$

Recalling $R_n(-\lambda) = (S_n + \lambda I_p)^{-1}$, an application of the Sherman–Morrison Formula yields that

$$R_n(-\lambda) = R_n^{(j)}(-\lambda) - \frac{R_n^{(j)}(-\lambda) \Sigma_p^{1/2} Y_j Y_j^T \Sigma_p^{1/2} R_n^{(j)}(-\lambda)}{1 + Y_j^T \Sigma_p^{1/2} R_n^{(j)}(-\lambda) \Sigma_p^{1/2} Y_j}.$$

The remainder of Appendix A is organized as follows. Several auxiliary lemmas are stated in Section A.1, the proof of Theorem 2.2 is worked out in Section A.2 and the proof of Theorem 3.1 is given in Section A.3.

A.1 Technical lemmas

Lemma A.1. Suppose that $Y \sim N(0, I_p)$ and A and B are $p \times p$ symmetric matrices. Then

$$\mathbb{E}[\text{tr}(AYY^T BYY^T)] = 2\text{tr}(AB) + \text{tr}(A)\text{tr}(B). \quad (\text{A.3})$$

The statement of the theorem follows from direct calculations. In what follows, let $\|\cdot\|$ be the operator norm of a matrix and $\|\cdot\|_F$ its Frobenius norm.

Lemma A.2. (Hanson–Wright inequality). Let $Y = (Y_1, \dots, Y_n)^T \in \mathbb{R}^n$ be a random vector with independent components Y_i having $\mathbb{E}[Y_i] = 0$ and uniformly bounded ψ_2 -norm (sub-Gaussian norm)

$$\|Y_i\|_{\psi_2} = \sup_{p \geq 1} \frac{1}{\sqrt{p}} (\mathbb{E}[|Y_i|^p])^{1/p} \leq K, \quad i = 1, \dots, n,$$

where $K > 0$ is a constant. Let A be an $n \times n$ matrix. Then, for any $t \geq 0$,

$$\mathbb{P}(|Y^T A Y - \mathbb{E}[Y^T A Y]| > t) \leq 2 \exp \left\{ -c \min \left(\frac{t^2}{K^4 \|A\|_F^2}, \frac{t}{K^2 \|A\|} \right) \right\},$$

where $c > 0$ is a constant.

Lemma A.3. Suppose we have a symmetric matrix A . Then for $\lambda > 0$ and any integer $k \geq 1$,

$$\left| \text{tr} \left[(R_n(-\lambda))^k A \right] - \text{tr} \left[(R_n^{(j)}(-\lambda))^k A \right] \right| \leq \frac{k \|A\|}{\lambda^k}.$$

The proof is given in the Supplementary Material document.

A.2 Proof of Theorem 2.2

Let c_1, c_2 and c_3 denote some universal positive constants, independent of λ . To lighten notation, some fixed parameters are ignored in the following expressions when it does not cause ambiguity; for example, weights π in $Q(\lambda, \gamma; \pi)$ may be dropped. It's first shown, in the following propositions, concentration of some quantities with proofs included in Supplementary Material document. Recall that $\hat{\phi} = p^{-1} \text{tr}(S_n)$ and $\phi = \int \tau dH(\tau)$.

Proposition A.1. If Assumption 2.1 is satisfied, then for any $t > 0$,

$$\mathbb{P} \left(\left| \hat{\phi} - \mathbb{E}\hat{\phi} \right| > t \right) \leq c_1 \exp \{ -\min(c_2 n t^2, c_3 n t) \}.$$

Moreover, $\sqrt{n} \left| \mathbb{E}\hat{\phi} - \phi \right| \rightarrow 0$, as $n \rightarrow \infty$, since $\mathbb{E}\hat{\phi} = \int \tau dH_p(\tau)$.

Proposition A.2. Define $m_{F_{n,p}}^{(k)}(-\lambda)$ to be the k -th order derivative of $m_{F_{n,p}}(-\lambda)$ and $m_F^{(k)}(-\lambda)$ to be the k th order derivative of $m_F(-\lambda)$. If Assumption 2.1 is satisfied, then for any $t > 0$, integer k and $\lambda \in [\underline{\lambda}, \bar{\lambda}]$,

$$\mathbb{P} \left(\left| m_{F_{n,p}}^{(k)}(-\lambda) - \mathbb{E} m_{F_{n,p}}^{(k)}(-\lambda) \right| > t \right) \leq c_1 \exp \{ -c_2 n t^2 \}.$$

Moreover,

$$\sqrt{n} \left| \mathbb{E} m_{F_{n,p}}^{(k)}(-\lambda) - m_F^{(k)}(-\lambda) \right| \rightarrow 0.$$

It follows, as continuous and monotone functions in λ ,

$$\sup_{\lambda \in [\underline{\lambda}, \bar{\lambda}]} \left| m_{F_{n,p}}^{(k)}(-\lambda) - m_F^{(k)}(-\lambda) \right| \xrightarrow{P} 0.$$

Proposition A.3. *If Assumption 2.1 is satisfied, then for any $\lambda \in [\underline{\lambda}, \bar{\lambda}]$,*

$$\frac{\partial}{\partial \lambda} \hat{\Theta}_1(\lambda, \gamma_n) = -\frac{1}{p} \text{tr} \left[(R_n(-\lambda))^2 \Sigma_p \right] + o_p(n^{-1/4}).$$

Proposition A.4. *If Assumption 2.1 is satisfied, then for any $\lambda \in [\underline{\lambda}, \bar{\lambda}]$,*

$$\frac{\partial^2}{\partial \lambda^2} \hat{\Theta}_1(\lambda, \gamma_n) = \frac{2}{p} \text{tr} \left[(R_n(-\lambda))^3 \Sigma_p \right] + o_p(1).$$

Proofs of Propositions A.1 and A.2 use sub-Gaussianity of the observations and the asymptotic behavior of linear spectral statistics of random Wishart matrices, respectively. Propositions A.3 and A.4 use a specific form of resolvent decomposition technique used in Chen et al. (2011). To save space, these proofs are given in the Supplementary Material.

Proof of (2.13) of Theorem 2.2. To show the existence of a sequence of local maximizers of $\hat{Q}_n(\lambda, \gamma_n)$ as stated, it suffices to show that for any $\varepsilon \in (0, 1)$, there exists a constant $K > 0$, and an integer n_ε , such that, for $t = Kn^{-1/4}$,

$$\mathbb{P} \left(\hat{Q}_n(\lambda_\infty \pm t, \gamma_n) - \hat{Q}_n(\lambda_\infty, \gamma_n) \leq 0 \right) \geq \varepsilon$$

for all $n \geq n_\varepsilon$. If we use a stochastic term $\delta(t)$ to measure the difference between $\hat{Q}_n(\lambda, \gamma_n)$ and $Q(\lambda, \gamma)$ at $\lambda = \lambda_\infty \pm t$ and λ_∞ , considering λ_∞ to be in the interior of $[\underline{\lambda}, \bar{\lambda}]$, a second-order Taylor expansion yields

$$\begin{aligned} \hat{Q}_n(\lambda_\infty \pm t, \gamma_n) - \hat{Q}_n(\lambda_\infty, \gamma_n) &= Q(\lambda_\infty \pm t, \gamma) - Q(\lambda_\infty, \gamma) + \delta(\pm t) \\ &= \frac{t^2}{2} \frac{\partial^2}{\partial \lambda^2} Q(\lambda_\infty, \gamma) + O(t^3) + \delta(\pm t) \end{aligned}$$

Since $O(t^3)$ is a smaller order term as $n \rightarrow \infty$ and $\frac{\partial^2}{\partial \lambda^2} Q(\lambda_\infty, \gamma) < 0$, it suffices to show that $\sqrt{n}|\delta(\pm t)| = O_p(1)$ with a uniform tail bound in t . Again by Taylor expansion,

$$\begin{aligned} \sqrt{n}\delta(\pm t) &= \sqrt{nt} \left\{ \frac{\partial}{\partial \lambda} \hat{Q}_n(\lambda_\infty, \gamma_n) - \frac{\partial}{\partial \lambda} Q(\lambda_\infty, \gamma) \right\} + \frac{\sqrt{nt}^2}{2} \left\{ \frac{\partial^2}{\partial \lambda^2} \hat{Q}_n(\lambda_\infty, \gamma_n) - \frac{\partial^2}{\partial \lambda^2} Q(\lambda_\infty, \gamma) \right\} \\ &\quad + \frac{\sqrt{nt}^3}{6} \frac{\partial^3}{\partial \lambda^3} \hat{Q}_n(\lambda_\infty + \alpha t, \gamma_n) - \frac{\sqrt{nt}^3}{6} \frac{\partial^3}{\partial \lambda^3} Q(\lambda_\infty + \alpha t, \gamma) \end{aligned}$$

for some $\alpha \in [0, 1]$. As continuous functions of $(m_F(-\lambda_\infty), m'_F(-\lambda_\infty), m_F^{(3)}(-\lambda_\infty), m_F^{(4)}(-\lambda_\infty), \phi, \gamma)$ and their empirical counterparts, by Proposition A.1–A.2,

$$\begin{aligned} n^{1/4} \left| \frac{\partial}{\partial \lambda} \hat{Q}_n(\lambda_\infty, \gamma) - \frac{\partial}{\partial \lambda} Q(\lambda_\infty, \gamma) \right| &\xrightarrow{P} 0 \\ \left| \frac{\partial^2}{\partial \lambda^2} \hat{Q}_n(\lambda_\infty, \gamma) - \frac{\partial^2}{\partial \lambda^2} Q(\lambda_\infty, \gamma) \right| &\xrightarrow{P} 0 \\ \sup_{\lambda \in [\underline{\lambda}, \bar{\lambda}]} \left| \frac{\partial^3}{\partial \lambda^3} \hat{Q}_n(\lambda, \gamma) \right| + \left| \frac{\partial^3}{\partial \lambda^3} Q(\lambda, \gamma) \right| &= O_p(1) \end{aligned}$$

which completes the proof. If λ_∞ is on the boundary and $\frac{\partial}{\partial \lambda} Q(\lambda_\infty, \gamma) < 0$, similar results follow from a first-order Taylor expansion. \square

Proof of (2.14) of Theorem 2.2. It remains to verify (2.14). To this end, note that it suffices to prove that

$$\begin{aligned} & \sqrt{p} \left| \frac{1}{p} \text{RHT}(\lambda_n) - \hat{\Theta}_1(\lambda_n, \gamma_n) - \frac{1}{p} \text{RHT}(\lambda_\infty) + \hat{\Theta}_1(\lambda_\infty, \gamma_n) \right| \\ & \leq \sqrt{p} \left| \frac{1}{p} \frac{\partial}{\partial \lambda} \text{RHT}(\lambda_\infty) - \frac{\partial}{\partial \lambda} \hat{\Theta}_1(\lambda_\infty, \gamma_n) \right| |\lambda_n - \lambda_\infty| \\ & + \frac{\sqrt{p}}{2} \left| \frac{1}{p} \frac{\partial^2}{\partial \lambda^2} \text{RHT}(\lambda_\infty) - \frac{\partial^2}{\partial \lambda^2} \hat{\Theta}_1(\lambda_\infty, \gamma_n) \right| |\lambda_n - \lambda_\infty|^2 \\ & + \frac{\sqrt{p}}{6} \left| \frac{1}{p} \frac{\partial^3}{\partial \lambda^3} \text{RHT}(\lambda^*) - \frac{\partial^3}{\partial \lambda^3} \hat{\Theta}_1(\lambda^*, \gamma_n) \right| |\lambda_n - \lambda_\infty|^3 \xrightarrow{P} 0 \end{aligned}$$

where λ^* is in between λ_∞ and λ_n . So it's enough to show

$$p^{1/4} \left| \frac{1}{p} \frac{\partial}{\partial \lambda} \text{RHT}(\lambda_\infty) - \frac{\partial}{\partial \lambda} \hat{\Theta}_1(\lambda_\infty, \gamma_n) \right| \xrightarrow{P} 0, \quad (\text{A.4})$$

$$\left| \frac{1}{p} \frac{\partial^2}{\partial \lambda^2} \text{RHT}(\lambda_\infty) - \frac{\partial^2}{\partial \lambda^2} \hat{\Theta}_1(\lambda_\infty, \gamma_n) \right| \xrightarrow{P} 0, \quad (\text{A.5})$$

$$\sup_{\lambda \in [\underline{\lambda}, \bar{\lambda}]} \left| \frac{1}{p} \frac{\partial^3}{\partial \lambda^3} \text{RHT}(\lambda) - \frac{\partial^3}{\partial \lambda^3} \hat{\Theta}_1(\lambda, \gamma_n) \right| = O_p(1). \quad (\text{A.6})$$

Next, $\mathbb{E} \left| p^{-1} \frac{\partial^3}{\partial \lambda^3} \text{RHT}(\lambda) \right| \leq \mathbb{E} \left| \frac{1}{p \lambda^4} \frac{n_1 n_2}{n_1 + n_2} (\bar{X}_1 - \bar{X}_2)^T (\bar{X}_1 - \bar{X}_2) \right| = O(1)$ for all $\lambda \in [\underline{\lambda}, \bar{\lambda}]$. And Proposition A.2 shows the convergence of $\frac{\partial^3}{\partial \lambda^3} \hat{\Theta}_1(\lambda, \gamma_n)$ to $\frac{\partial^3}{\partial \lambda^3} \Theta_1(\lambda, \gamma)$ uniformly on $\lambda \in [\underline{\lambda}, \bar{\lambda}]$. (A.6) holds.

For proving (A.4) and (A.5), note that Propositions A.3 and A.4 showed the convergence of $\frac{\partial}{\partial \lambda} \hat{\Theta}_1(\lambda, \gamma_n)$ to $-\frac{1}{p} \text{tr} [(R_n(-\lambda))^2 \Sigma_p]$, and the convergence of $\frac{\partial^2}{\partial \lambda^2} \hat{\Theta}_1(\lambda, \gamma_n)$ to $\frac{2}{p} \text{tr} [(R_n(-\lambda))^3 \Sigma_p]$. So the proof will be complete if we can show

$$p^{1/4} \left| \frac{1}{p} \frac{\partial}{\partial \lambda} \text{RHT}(\lambda_\infty) + \frac{1}{p} \text{tr} [(R_n(-\lambda_\infty))^2 \Sigma_p] \right| \xrightarrow{P} 0, \quad (\text{A.7})$$

$$\left| \frac{1}{p} \frac{\partial^2}{\partial \lambda^2} \text{RHT}(\lambda_\infty) - \frac{2}{p} \text{tr} [(R_n(-\lambda_\infty))^3 \Sigma_p] \right| \xrightarrow{P} 0. \quad (\text{A.8})$$

It is worth mentioning that all arguments in the proof up to now do not require normality assumption of observations. The proofs of (A.7) and (A.8) would be almost free if we have Gaussianity. To see it, note first

$$\begin{aligned} \mathbb{E} \frac{1}{p} \frac{\partial}{\partial \lambda} \text{RHT}(\lambda_\infty) &= -\mathbb{E} \frac{1}{p} \text{tr} [(R_n(-\lambda_\infty))^2 \Sigma_p] \\ \mathbb{E} \frac{1}{p} \frac{\partial^2}{\partial \lambda^2} \text{RHT}(\lambda_\infty) &= 2\mathbb{E} \frac{1}{p} \text{tr} [(R_n(-\lambda_\infty))^3 \Sigma_p] \end{aligned} \quad (\text{A.9})$$

From Lemma A.2,

$$\begin{aligned} \frac{1}{p} \frac{\partial}{\partial \lambda} \text{RHT}(\lambda_\infty) &= \mathbb{E} \frac{1}{p} \frac{\partial}{\partial \lambda} \text{RHT}(\lambda_\infty) + O_p\left(\frac{1}{\sqrt{p}}\right), \\ \frac{1}{p} \frac{\partial^2}{\partial \lambda^2} \text{RHT}(\lambda_\infty) &= \mathbb{E} \frac{1}{p} \frac{\partial^2}{\partial \lambda^2} \text{RHT}(\lambda_\infty) + O_p\left(\frac{1}{\sqrt{p}}\right). \end{aligned} \quad (\text{A.10})$$

With a difference bound shown in Lemma A.3, applying McDiarmid inequality, we have

$$\frac{1}{p} \text{tr} [(R_n(-\lambda))^2 \Sigma_p] = \mathbb{E} \frac{1}{p} \text{tr} [(R_n(-\lambda))^2 \Sigma_p] + O_p\left(\frac{1}{\sqrt{p}}\right),$$

$$\frac{1}{p} \text{tr} [(R_n(-\lambda))^3 \Sigma_p] = \mathbb{E} \frac{1}{p} \text{tr} [(R_n(-\lambda))^3 \Sigma_p] + O_p\left(\frac{1}{\sqrt{p}}\right),$$

which completes the proof under normality assumption.

Under non-Gaussianity, although (A.9) and (A.10) would fail, we can still show (A.7) and (A.8) which will complete all technical support of the theorem. However, the proof is long and hence the details are included in the Supplementary Material. \square

A.3 Proof of Theorem 3.1

To prove the process convergence stated in Theorem 3.1, convergence of finite-dimensional distributions and tightness need to be verified. The joint asymptotic normality of $(T_{n,p}(\lambda_1), \dots, T_{n,p}(\lambda_k))$ for a selection of k regularization parameters under Gaussianity follows from a small extension of the arguments used in Chen et al. (2011). In the Supplementary Material, the statement is more generally proved for observations belonging to $\mathcal{L}(c_1, c_2)$, so that the Gaussian proof is omitted here. The remaining parts are then to show tightness and to compute the asymptotic covariance kernel Γ in (3.2).

Proof of Theorem 3.1. (a) The convergence of the finite-dimensional distributions may be found in the Supplement Material.

(b) To show tightness, note first that Proposition A.2 yields $\hat{\Theta}_2(\lambda, \gamma_n) \rightarrow_p \Theta_2(\lambda, \gamma)$ uniformly on $[\underline{\lambda}, \bar{\lambda}]$. This implies tightness of $(\hat{\Theta}_2(\lambda, \gamma_n) : \lambda \in [\underline{\lambda}, \bar{\lambda}])$. The sequence $\sqrt{n}(\frac{1}{p} \text{RHT}(\lambda) - \hat{\Theta}_1(\lambda, \gamma_n))$ is shown to be tight in Pan & Zhou (2011, Section 4) more generally for observations with finite fourth moments. Although their arguments are in a one-sample testing framework, they can easily be generalized to the two-sample testing case. Together with $\inf_{\lambda \in [\underline{\lambda}, \bar{\lambda}]} \Theta_2(\lambda, \gamma) > 0$, the convergence of the process follows.

(c) To compute the limiting covariance kernel, let $\lambda \neq \lambda' \in \mathbb{R}^+$. Direct calculations and an application of Lemma A.1 show that

$$\begin{aligned} \Gamma(\lambda, \lambda') &= \frac{1}{2\sqrt{\Theta_2(\lambda, \gamma)\Theta_2(\lambda', \gamma)}} \lim_{n \rightarrow \infty} \frac{1}{p} \text{Cov}[\text{RHT}(\lambda), \text{RHT}(\lambda') | S_n] \\ &= \frac{1}{\sqrt{\Theta_2(\lambda, \gamma)\Theta_2(\lambda', \gamma)}} \lim_{n \rightarrow \infty} \frac{1}{p} \text{tr}[R_n(-\lambda) \Sigma_p R_n(-\lambda') \Sigma_p]. \end{aligned} \quad (\text{A.11})$$

The proof of this result in the non-Gaussian case is considerably harder and is included in the Supplementary Material.

Use of the identity $I_p - \lambda R_n(-\lambda) = R_n(-\lambda) S_n$, and the Sherman–Morrison rank one updating formula for matrix inverses yield

$$\begin{aligned} &R_n(-\lambda') - \lambda R_n(-\lambda) R_n(-\lambda') \\ &= \sum_{j=1}^{n-2} \frac{R_n^{(j)}(-\lambda) \Sigma_p^{1/2} Y_j Y_j^T \Sigma_p^{1/2} R_n(-\lambda')}{1 + Y_j^T \Sigma_p^{1/2} R_n^{(j)}(-\lambda) \Sigma_p^{1/2} Y_j} \\ &= \sum_{j=1}^{n-2} \frac{R_n^{(j)}(-\lambda) \Sigma_p^{1/2} Y_j Y_j^T \Sigma_p^{1/2} R_n^{(j)}(-\lambda')}{\{1 + Y_j^T \Sigma_p^{1/2} R_n^{(j)}(-\lambda) \Sigma_p^{1/2} Y_j\} \{1 + Y_j^T \Sigma_p^{1/2} R_n^{(j)}(-\lambda') \Sigma_p^{1/2} Y_j\}}. \end{aligned}$$

This implies that

$$\begin{aligned}
& \frac{1}{p} \text{tr}[R_n(-\lambda')\Sigma_p] - \lambda \frac{1}{p} \text{tr}[R_n(-\lambda)R_n(-\lambda')\Sigma_p] \\
&= \frac{1}{p} \sum_{j=1}^{n-2} \frac{Y_j^T \Sigma_p^{1/2} R_n^{(j)}(-\lambda') \Sigma_p R_n^{(j)}(-\lambda) \Sigma_p^{1/2} Y_j}{\{1 + Y_j^T \Sigma_p^{1/2} R_n^{(j)}(-\lambda) \Sigma_p^{1/2} Y_j\} \{1 + Y_j^T \Sigma_p^{1/2} R_n^{(j)}(-\lambda') \Sigma_p^{1/2} Y_j\}} \\
&= \frac{\frac{1}{p} \text{tr}[R_n(-\lambda) \Sigma_p R_n(-\lambda') \Sigma_p]}{\{1 + \gamma_n \frac{1}{p} \text{tr}[R_n(-\lambda) \Sigma_p]\} \{1 + \gamma_n \frac{1}{p} \text{tr}[R_n(-\lambda') \Sigma_p]\}} + \zeta_5. \tag{A.12}
\end{aligned}$$

In the last step, ζ_5 is a remainder term, which can be shown to be $o_p(1)$ along the same lines as ζ_3 and ζ_4 .

Now, using the fact that

$$(\lambda' - \lambda) \text{tr}[R_n(-\lambda)R_n(-\lambda')\Sigma_p] = \text{tr}[R_n(-\lambda)\Sigma_p] - \text{tr}[R_n(-\lambda')\Sigma_p]$$

it follows that

$$\begin{aligned}
& \frac{1}{p} \text{tr}[R_n(-\lambda) \Sigma_p R_n(-\lambda') \Sigma_p] \\
&= \{1 + \frac{\gamma_n}{p} \text{tr}[R_n(-\lambda) \Sigma_p]\} \{1 + \frac{\gamma_n}{p} \text{tr}[R_n(-\lambda') \Sigma_p]\} \left(\frac{\lambda' \text{tr}[R_n(-\lambda') \Sigma_p] - \lambda \text{tr}[R_n(-\lambda) \Sigma_p]}{p(\lambda' - \lambda)} \right) + o_p(1) \\
&= \{1 + \gamma \Theta_1(\lambda, \gamma_n)\} \{1 + \gamma \Theta_1(\lambda', \gamma_n)\} \left(\frac{\lambda' \Theta_1(\lambda', \gamma_n) - \lambda \Theta_1(\lambda, \gamma_n)}{\lambda' - \lambda} \right) + o_p(1). \tag{A.13}
\end{aligned}$$

The last step uses that, under Assumption 2.1, the asymptotic relation (2.7) holds. Combining (A.11)–(A.13) leads to expression (3.2). The proof is complete. \square

References

- Anderson, T.W. (1984). *An Introduction to Multivariate Statistical Analysis, 2nd Edition*. Wiley, New York.
- Bai, Z.D. and Saranadasa, H. (1996). Effect of high dimension: by an example of a two sample problem. *Statistica Sinica* **6**, 311–329.
- Bai, Z. and Silverstein, J.W. (1998). No eigenvalues outside the support of the limiting spectral distribution of large-dimensional sample covariance matrices. *Annals of Probability* **26**, 316–345.
- Bai, Z. and Silverstein, J.W. (2004). CLT for linear spectral statistics of large-dimensional sample covariance matrices. *Annals of Probability* **32** 553–605.
- Bai, Z. and Silverstein, J.W. (2010). *Spectral Analysis of Large Dimensional Random Matrices*. Springer-Verlag, New York.
- Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B* **57**, 289–300.

- Bergamaschi, A., Kim, Y. H., Wang, P., Sørbye, T., Hernandez-Boussard, T., Lonning, P. E., Tibshirani, R., Børresen-Dale, A. L. and Pollack, J. R. (2006). Distinct patterns of DNA copy number alteration are associated with different clinicopathological features and gene-expression subtypes of breast cancer. *Genes Chromosomes Cancer* **45**(11), 1033–40.
- Biswas, M. and Ghosh, A.K. (2014). A nonparametric two-sample test applicable to high dimensional data. *Journal of Multivariate Analysis* **123**, 160–171.
- Cai, T.T., Liu, W. and Xia, Y. (2014). Two-sample test of high dimensional means under dependence. *Journal of the Royal Statistical Society, Series B* **76**, 349–372.
- Cancer Genome Atlas Network (2012). Comprehensive molecular portraits of human breast tumors. *Nature* **490**(7418), 61–70.
- Chang, J., Zhou, W. and Zhou, W. X. (2014) Simulation-based hypothesis testing of high dimensional means under covariance heterogeneity – an alternative road to high dimensional tests. *arXiv:1406.1939*.
- Chatterjee, S. (2009). Fluctuations of eigenvalues and second order Poincaré inequalities. *Probability Theory and Related Fields* **143**, 1–40.
- Chen, S.X. and Qin, Y.-L. (2010). A two-sample test for high-dimensional data with applications to gene-set testing. *The Annals of Statistics* **38**, 808–835.
- Chen, L., Paul, D., Prentice, R. L. and Wang, P. (2011). A regularized Hotelling’s T^2 test for pathway analysis in proteomic studies. *Journal of the American Statistical Association* **106**, 1345–1360.
- Chen, S. X., Li, J. and Zhong, P. (2014) Two-sample tests for high dimensional means with thresholding and data transformation. *arXiv:1410.2848*.
- Creighton, C. J. (2012). The molecular profile of luminal B breast cancer. *Biologics: Targets & Therapy* **6**, 289.
- Dempster, A. P. (1958). A high dimensional two sample significance test. *The Annals of Mathematical Statistics* **29**, 995–1010.
- Dempster, A. P. (1960). A significance test for the separation of two highly multivariate small samples. *Biometrics* **16**, 41–50.
- Dong, K., Pang, H., Tong, T. and Genton, M. G. (2016). Shrinkage-based diagonal Hotelling’s tests for high-dimensional small sample size data. *Journal of Multivariate Analysis* **143**, 127–142.
- Ellis, M. J., Gillette, M., Carr, S. A., Paulovich, A. G., Smith, R. D., Rodland, K. K., Townsend, R. R., Kinsinger, C., Mesri, M., Rodriguez, H., Liebler, D. C. and Clinical Proteomic Tumor Analysis Consortium (CPTAC) (2013). Connecting genomic alterations to cancer biology with proteomics: The NCI Clinical Proteomic Tumor Analysis Consortium. *Cancer Discovery* **3**(10), 1108–1112.

- Gregory, K. B., Carrol, R. J., Baladandayuthapani, V. and Lahiri, S. N. (2015). A two-sample test for equality of means in high dimension. *Journal of the American Statistical Association* **110**, 837–849.
- Guo, B. and Chen, S. X. (2016). Tests for high dimensional generalized linear models. *Journal of the Royal Statistical Society, Series B*, to appear. doi: 10.1111/rssb.12152.
- Jiang, J., Li, C., Paul, D., Yang, C. and Zhao H. (2016). On high dimensional misspecified mixed model analysis in genome-wide association studies. *arXiv Preprint arXiv:1404.2355*
- Lamy, P.-J., Fina, F., Bascoul–Molle, C., Laberrenne, A.-C., Martin, P.-M., Ouafik, L., Jacot, W., and others (2011). Quantification and clinical relevance of gene amplification at chromosome 17q12-q21 in human epidermal growth factor receptor 2-amplified breast cancers. *Breast Cancer Research* **13**, R15.
- Liu, H., Aue, A. and Paul, D. (2015). On the Marčenko–Pastur law for linear time series. *The Annals of Statistics* **43**, 675–712.
- Lopes, M. E., Jacob, L. and Wainwright, M. J. (2011). A more powerful two-sample test in high dimensions using random projection. *Advances in Neural Information Processing Systems (NIPS)*, 2011.
- Lugosi, G., Massart, P. and Boucheron, S. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press.
- Mertins, P., Mani, D. R., Ruggles, K. V., Gillette, M. A., Clauser, K. R., Wang, P., Wang, X., Qiao, J. W., Cao, S. and Petralia, F. and others (2016). Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature* **534**(7605), 55–62.
- Muirhead, R.J. (1982). *Aspects of Multivariate Statistical Theory*. Wiley, New York.
- Pan, G. M. and Zhou, W. (2011). Central limit theorem for Hotelling’s T^2 statistic under large dimension. *The Annals of Applied Probability* **21**, 1860–1910.
- Park, J. and Ayyala, D. N. (2013). A test for the mean vector in large dimension and small samples. *Journal of Statistical Planning and Inference* **143**(5), 929–943.
- Paul, D. and Aue, A. (2014). Random matrix theory in statistics: A review. *Journal of Statistical Planning and Inference* **150**, 1–29.
- Paulovich, A. G., Billheimer, D., Ham, A. J., Vega-Montoto, L., Rudnick, P. A., Tabb, D. L., Wang, P., Blackman, R. K., Bunk, D. M., Cardasis, H. L., Clauser, K. R., Kinsinger, C. R., Schilling, B., Tegeler, T. J., Variyath, A. M., Wang, M., Whiteaker, J. R., Zimmerman, L. J., Fenyo, D., Carr, S. A., Fisher, S. J., Gibson, B. W., Mesri, M., Neubert, T. A., Regnier, F. E., Rodriguez, H., Spiegelman, C., Stein, S. E., Tempst, P. and Liebler D. C. (2010). Interlaboratory study characterizing a yeast performance standard for benchmarking LC-MS platform performance. *Molecular & Cellular Proteomics* **9**(2), 242–254.
- Srivastava, M. and Du, M. (2008). A test for the mean vector with fewer observations than the dimension. *Journal of Multivariate Analysis* **99**, 386–402.

- Srivastava, M. (2009). A test for the mean vector with fewer observations than the dimension under non-normality. *Journal of Multivariate Analysis* **100**, 518–532.
- Srivastava, M. S., Katayama, S. and Kano, Y. (2013). A two sample test in high dimensional data. *Journal of Multivariate Analysis* **114** 349–358.
- Srivastava, R., Li, P. and Ruppert, D. (2016). RAPTT: An exact two-sample test in high dimensions using random projections. *Journal of Computational and Graphical Statistics* (to appear).
- Tran, B. and Bedard, P. (2011). Luminal-B breast cancer and novel therapeutic targets. *Breast Cancer Research* **13**, 221.
- Vershynin, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In: *Eldar Y. and Kutyniok, G. (eds.). Compressed Sensing: Theory and Applications*, Cambridge University Press, Cambridge, pp. 210–268.
- Wang, R., Peng, L. and Qi, Y. (2015). Jackknife empirical likelihood test for equality of two high dimensional means. *Statistica Sinica* **23**, 667–690.
- Wang, L., Peng, B. and Li, R. (2015). A high-dimensional nonparametric multivariate test for mean vector. *Journal of the American Statistical Association* **110**, 1658–1669.